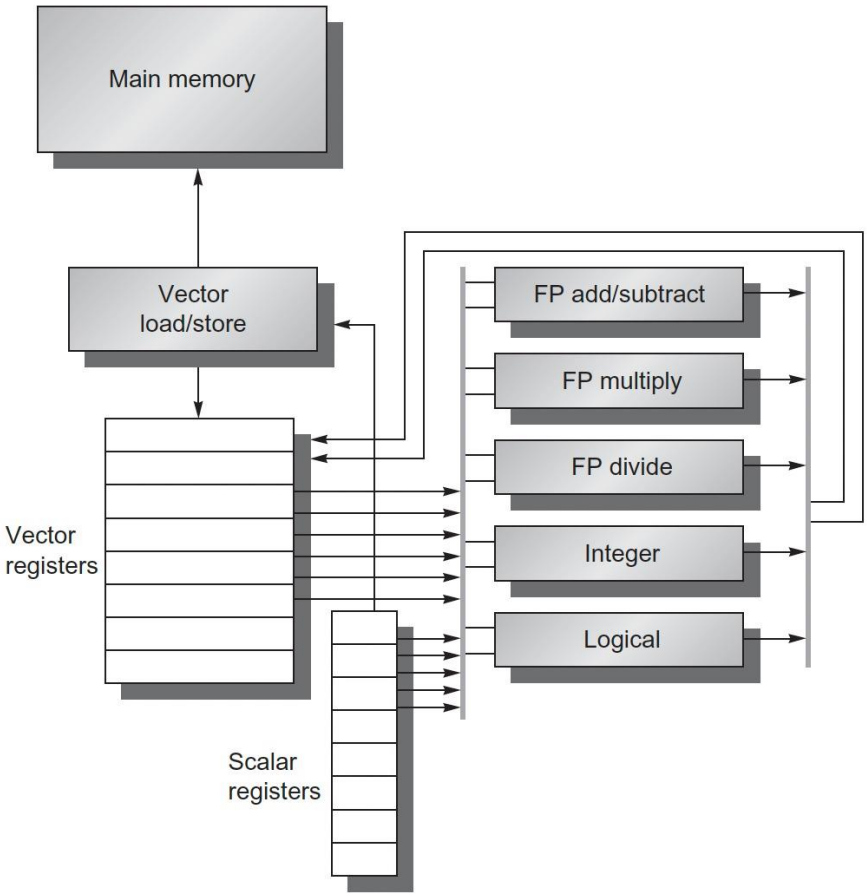
EEL 6764 001: Graduate Computer Architecture
Spring 2025
Instructor: Dr. Srinivas Katkoori
Homework 4
Solutions

1) **(15 pts) Vector Processing What is the basic idea of a vector processor?** The main idea behind vector processing is performing a single operation on multiple data at the same time. So, instead of processing each data element individually, the elements are placed into vectors/ arrays for them to be processed in "bulk". In other words, vector processing allows computers to go from sequential data processing (one operation for each data element) to matrix or array processing (one operation for the whole array or matrix). Vector processing is also known as SIMD (Single Instruction Multiple Data)
2) **(20 pts) Vector Processing Draw the block diagram of RISC-V Vector architecture. Briefly explain the various blocks in the architecture ?**



Block Diagram of RISC-V Vector Architecture

- Each register holds a 64-element, 64 bits/element vector. Register file has 16 read ports and 8 write ports. Other special purpose registers, i.e. VLR/MVL .
- Fully pipelined – start a new op on every cycle, Control unit detects data & structural hazards.
- Memory needs to have high bandwidth to support vector load/store.
- Fully pipelined – one word to/from memory per cycle after initial latency, Also handle scalar load/store.
- 32 GP registers and 32 FP registers.

3) (20 pts) Vector Processing With an example, explain the basic idea of the Following techniques:

a. Vector chaining.
b. Strip Mining
c. Vector Stride
d. Gather and Scatter

a) Vector chaining: The main idea of vector chaining is the optimization of instruction execution (vector version of forwarding). The next vector operation starts when the result from previous operation on the first element is finished.

Example:

LV v1
MULV v3, v1, v2
ADDV v5 v3, v4

b) Strip Mining: Strip-mining, or loop sectioning, is a technique for manipulating loops such that the loop is partitioned in smaller segments or "strips". It may improve memory performance. Reduces the loop's size by a factor of the length of the vector maximum length.

Example:

This vector loop:

```
for (i=0; i < n; i = i + 1)
Y[i] = a * X[i] + Y[i]
```

is converted to the following:

```
low = 0;
VL = (n % MVL);
for(j=0; j <= (n/MVL); j=j+1) {
for(i=low; i <(low + VL); i=i+1)
Y[i] = a * X[i] + Y[i];
low = low + VL;
VL = MVL
}
```

c) Vector Stride: Technique for handling multi-dimensional arrays. Stride is the distance that separates elements in an array to be gathered into a single register. Vector base address and stride is stored in GP registers.

Example:

```
for(i=0; i<100; i=i+1)
for(j=0; j<100; j=j+1) {
A[i][j] = 0.0;
for(k=0; k < 100; k=k+1)
A[i][j] = A[i][j] + B[i][k]*D[k][j];
}
```

If D is stored in row-major layout then the stride is 100 double words.

d) Gather and Scatter: A way of addressing and improving memory access in sparse matrices (arrays which most elements are zero). It basically, creates a new matrix (index matrix) with the index of the elements that are non-zero and then maps it out to a new matrix (dense matrix).

Example:

4) (20 pts) Snooping Protocol Problem 5.1 parts (a), (b), (c) (Pages 446-447)

Cx.y is cache line y in core x.

a. C0: R AC20 ! C0.0: (S, AC20, 0020), returns 0020

b. C0: W AC20 80 ! C0.0: (M, AC20, 0080)

C3.0: (I, AC20, 0020)

c. C3: W AC20 80 ! C3.0: (M, AC20, 0080)

5) (15 pts) Coherence Protocols Explain how the following protocols work:

a. Snooping coherence protocol

Every cache track sharing status of each cache block

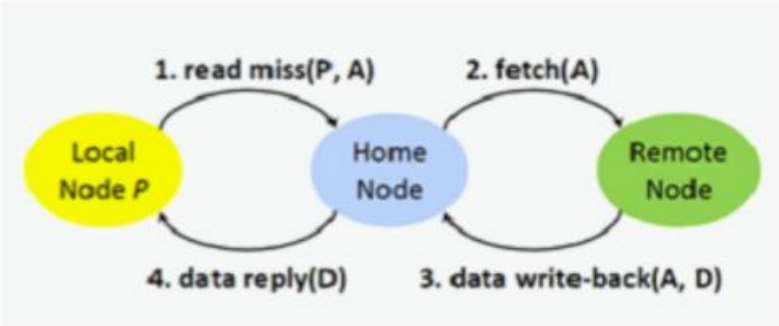
- Mem requests are broadcast on a bus to all cache
- Writes serialized naturally by a single bus

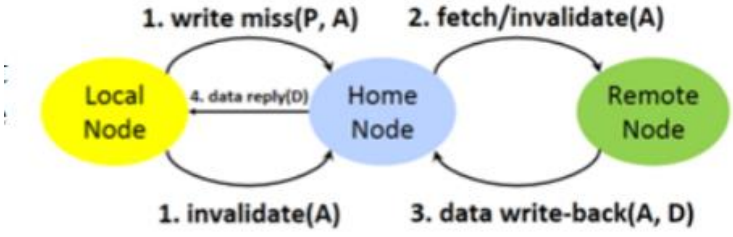
Used in shared memory multiprocessor systems. Each cache controller monitors, or "snoops", the bus that interconnects the processors and caches to determine if any other processors or caches have accessed the same memory block.

b. Directory based coherence protocol.
The memory is distributed across the processors. Directory maintains block states and sends invalidation messages, i.e., tracks states of all local memory blocks. For each memory block, shared, uncached (invalid), and modified (exclusive)

6) (10 pts) Directory Protocol

a. Using a state-diagram like figure, show how a read miss from local node is handled by the home node and remote node. Briefly explain how it works.
Read miss: When a local node has a read miss, it will request the home node for a copy (arc 1). Home node then requests the latest copy from remote node (arc 2).
The remote node does the write-back to home node (arc3). Finally, the home node sends the latest copy to the requesting local node (arc 4).





Operation 2: invalidate, or fetch & Invalidate

b. Using a state-diagram like figure, show how a write miss from local node is handled by the home node and remote node. Briefly explain how it works.
Write miss: When a local node has a write miss, it will request the home node for a copy (arc 1) as well as invalidates the copy. The home node then requests the latest copy

from remote node (arc 2) as well as invalidates the remote nodes' copy. The remote node does the write back to home node (arc 3). Finally, the home node sends the latest copy to the requesting local node (arc 4).